Harrigan et al. (2017) first response to reviewers - Reviewer 2: RC2 (Guillaume Thirel)

All reviewer 2 (Guillaume Thirel') comments are labelled consecutively, for example, comment 1 is R#2-1, with our responses to his comments given in blue text.

R#2-1. This manuscript presents an evaluation of ESP over the UK. The ensemble forecasts are based on the lumped conceptual GR4J model and past P and PET observations that were resampled as used as input to GR4J. These forecasts are compared to proxy observations (GR4J streamflows using P and PET observations) and a benchmark (resampling of these GR4J streamflows).

This paper is generally well written, very clear, and it makes a significant contribution to the HESS journal. However, I of course have some remarks that would deserve some attention from the authors, some of them not being minor. I am convinced that the authors will be able to handle that efficiently and allow the paper to be published.

We thank Guillaume Thirel very much for his supportive comments and constructive feedback that has helped us refine our paper, particularly his insights on hydrological modelling components.

Major comments:

R#2-2. The way ESP is thought of in this manuscript is a bit old fashioned in my opinion. It is true that first ESPs were using IHCs and past data, but this is not really the standard nowadays. Indeed, the standard is more what is called in the article NWS ESP. These forecasts are now a well-established method and are the reference, especially up to a month of lead time. I would advise the authors using a more modern terminology in the abstract and article or at least being more specific. Moreover, the justification of the choice of this method should be given.

We fully recognise that ESP, in its traditional form as used here, is a very simple method, and that alternative more-sophisticated ensemble hydrological forecasting techniques are becoming used more and more. We believe, however, there is still a need for benchmarking the skill of such simpler methods as traditional ESP is still considered a good alternative forecasting technique, in the absence of for example expensive seasonal climate forecasts. The choice of evaluating the forecast performance of a simple method like traditional ESP was motivated for three main reasons: 1) to provide a benchmark against which more complex methods could be evaluated for a range of lead times, up to 365 days - this is rarely done (nor possible with more computationally expensive techniques); 2) to identify when/where traditional ESP does not contain sufficient information to generate a skilful hydrological forecast, and henceforth where more complex methods, including use of dynamic atmospheric forecasts, are therefore essential for generating skilful hydrological forecasts; and 3) to formalise the skill of the hydrological seasonal forecasting systems currently used operationally in the UK (within the Hydrological Outlooks UK: http://www.hydoutuk.net/), through a national-scale analysis – the first time this has been done.

We will however edit the revised manuscript to:

a.) more clearly distinguish that it is ESP in its traditional form we are assessing. For example we will insert the following text in Pg2; L15: “In the traditional formulation of ESP as used in this paper,…” & Pg2 21: “Traditional ESP, while simple, is still widely used today in operational seasonal hydrological forecasting (e.g. US NWS and HOUK) and as a low cost forecast against which to benchmark potential skill improvements from more sophisticated hydro-meteorological ensemble prediction systems”.
b.) we will also provide stronger justification why this simple method is still used by many other today and indeed why we are examining it within this manuscript on Pg3; L5: “The previous studies demonstrate that

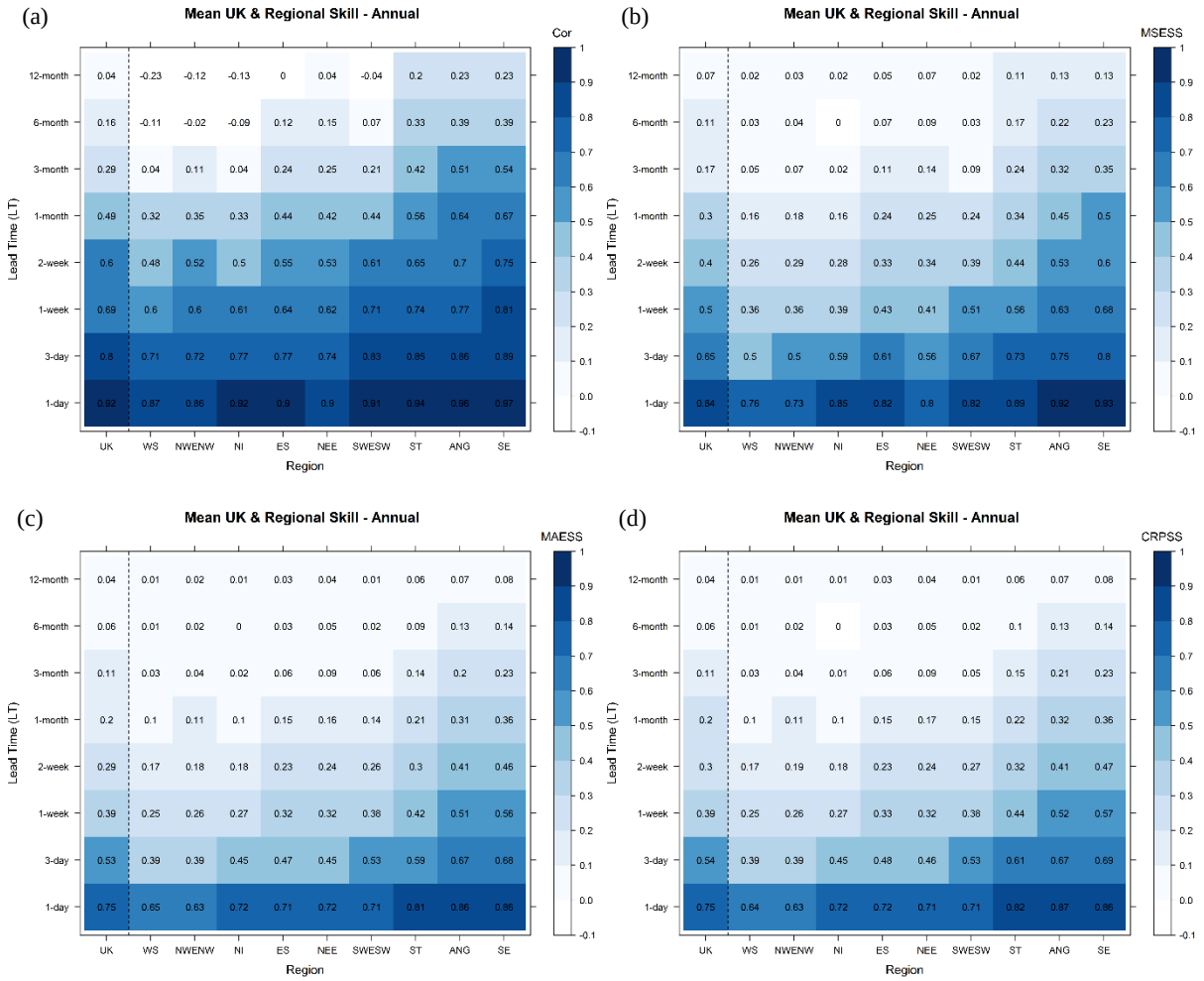
skilful forecasts can be made using the traditional ESP method at both short and long lead times in many regions around the world and given its relative ease of application and low computational cost remains a valuable ensemble hydrological forecasting approach. Although ESP is being used operationally within the UK, its skill has not yet been investigated at the catchment-scale within a rigorous hindcast experiment and is therefore the focus of this paper".

R#2-3. IHCs influence is high for short lead times and low for large lead times. Following the authors' sentence (P. 8, L. 2-4) that would mean that for short lead times, MSESS and CRPSS should be closer than for long lead times. However, we don't see that on Fig. 4, all lead times seem to have a similar difference between both SSs.

This comment and the comments from reviewer #3 sparked our curiosity of the impact of using different skill score metrics. We agree with comment R#3-6 that comparing MSESS (as the deterministic measure of ensemble mean) and CRPSS (as the probabilistic measure of full ensemble) in the way we have done in Figure 3 (and on Pg; L2-4 that you are referring to) is misleading as these two scores are not directly comparable. As reviewer #3 points out it is the Mean Absolute Error Skill Score (MAESS) that equals CRPSS for a deterministic forecast (also mentioned in the paper you recommend by Trinh et al., 2013), and would have been better to use instead of MSESS.

We have taken this suggestion on board and have tested four of the most common used metrics for assessing hydrological forecasts: Pearson's correlation coefficient (not a skill score: x = ensemble mean, y = proxy obs), MSESS (deterministic), MAESS (deterministic), and the CRPSS (probabilistic). Results from this analysis show that scores from the MAESS and CRPSS are very similar (see figure S2 below), and that the there is virtually no difference between the skill ensemble mean and full ensemble across lead times or regions (Figure S2 c and d). The results for correlation (Figure S2a) and MSESS (Figure S2b and same as Figure 6 in the original manuscript) are systematically higher than MAESS and CRPSS, not due to IHC influence etc. but simply due to the different formulation of these metrics. Their values on a 0 to 1 scale are not directly comparable. However, it must be made clear that it is only the ***magnitude*** of values that is different – the results/interpretation of ESP skill remain the same no matter which metric is used (so most/least skilful region, skill across initialisation months etc.).

We have decided to take the advice of review #3 wrt to skill scores and we will concentrate on CRPSS, as ESP is a probabilistic method. Given results are so similar between the full ensemble and deterministic ESP forecasts using MAESS, in the revised manuscript we will only use CRPSS (instead of MSESS) in Figures 3, 4, 5, 6, 7, and 8. Therefore, the text in Pg8; L2-4 referring to your will be modified accordingly. We think it's important to include the results of the comparison of the four scores and will include in as **supplementary Figure S2**.



Supplementary Figure 2: Heatmap of mean ESP skill across all 12 forecast initialisation months for the UK and for each of the nine hydroclimate regions ordered from least to most skilful (horizontal axis) at eight sample lead times (vertical axis). Skill is given by the a.) Pearson correlation coefficient (Cor.), b.) Mean Squared Error Skill Score (MSESS), c.) Mean Absolute Error Skill Score (MAESS), and d.) Continuous Ranked Probability Skill Score (CRPSS). Darker (lighter) shades showing higher (lower) skill; individual mean skill values are shown within each cell.

R#2-4. Section 4.1.2: this analysis is interesting. However, there is a second possible entry, in addition to the initialisation month, to take into account in my opinion: the lead time month. Indeed, some periods of the period are easier to predict (typically in between seasons are more prone to changing weather, which is difficult to predict sometimes); that may reflect on the scores, and could explain the differences that are highlighted here. Moreover, some scores can be impacted, for instances, by the streamflow characteristics. It is known that Nash-Sutcliffe (not used here) is higher for rivers with strong seasonality, or that CRPS is impacted by the streamflow magnitude (Trinh et al., 2013). I'm wondering to which extent the seasonal analysis (but also the spatial analysis actually!) can be impacted by such issues.

Thanks for these insights and references. First, the issue with CRPS being impacted by streamflow magnitude (as shown in Trinh et al., 2013) is not a problem in our analysis as we are using the CRPS skill score (CRPSS) so is not dependent on streamflow magnitude. However, the other issues could certainly be playing a minor or major role. As explained in R#2-1 the main aim of this work was to perform the first assessment of ESP skill over a range of lead times at the national scale. In order to identify future possible research avenues, we looked if any simple spatial/temporal patterns emerged from the analysis (i.e. Sections 4.1 and 4.2). The attribution of skill (the 'why' in Section 4.3) is meant as a first assessment of the apparent strong relationship between catchment storage and ESP skill. However, we have a discussion point attribution of different ESP

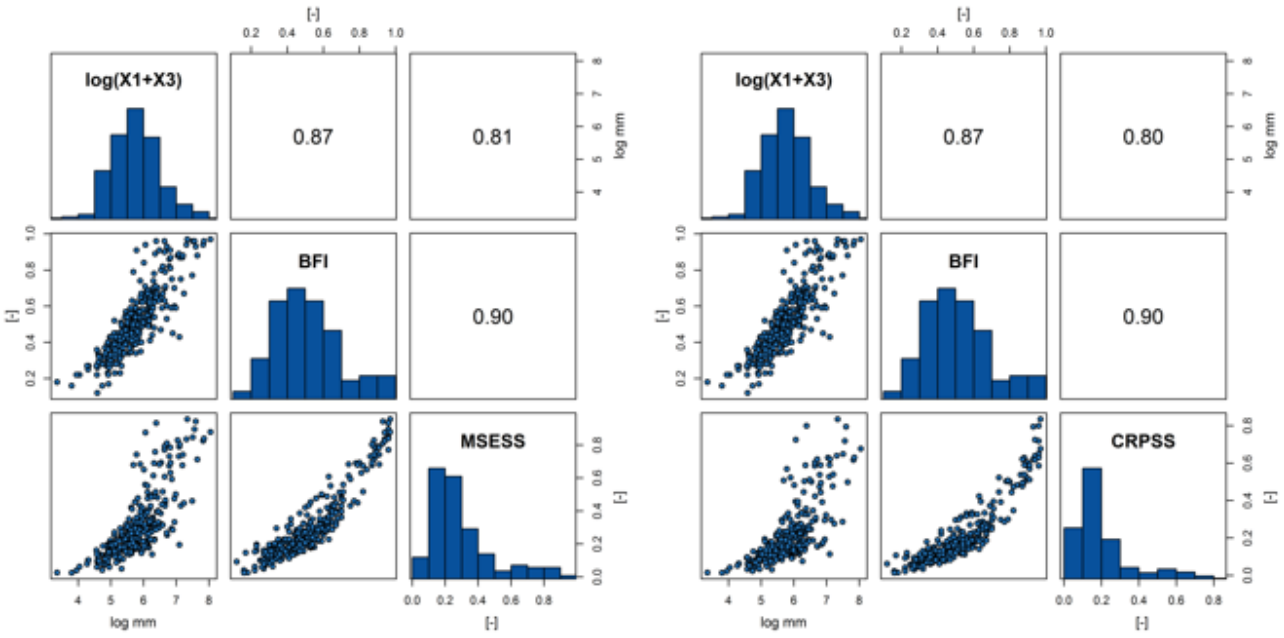
skills in transition monthly on Pg 10; L30-32:"Factors that might contribute to lower skilled forecasts initialised in spring include potentially higher uncertainty in IHC storage states and larger variability in rainfall across the forecast window (i.e. from late spring to early autumn). Further work should endeavour to attribute the lack of skill during transition seasons but this is outside the scope of this paper".

While we believe a full diagnostic and attribution assessment of the factors responsible for different ESP skills initialised in different times of the year is outside the scope of this paper, as it would require a much more detailed analysis over a complex range of issues, which would lengthen the paper considerably. We will make this much more clear in the revised manuscript and expand the discussion point on Pg10; L30-32 to a wider set of possible explanations of different hydrological forecast performance (e.g. influence of groundwater response, more variable weather conditions over the forecast period, and as GR4J is calibrated using all months there could be some interaction with better/worse simulation of IHCs for different months).

R#2-5. P. 9, L. 21-22: X1 is the production store capacity, and X3 the routing store capacity. It seems difficult to actually link them directly and specifically to soil and groundwater. However, their sum can be considered of the maximum amount of water in the basin (excluding the water in the river and snowpack) and as such it could be of interest including it in Fig. 8.

We agree that it is very difficult directly link X1 and X3 to soil moisture and groundwater, respectively. However, what is really of interest in this first assessment is the more general question of whether catchment storage is in any way related to ESP performance. We therefore will remove specific reference to linking skill directly to individual soil moisture/groundwater storage capacity model parameter values in the revised manuscript, but instead use your suggestion of viewing (X1 + X3) as total catchment storage capacity (minus snow and water in the river channel). E.g. Research Question 3 on Pg 3; L19 will be changed from "Where is ESP skilful, in terms of individual catchment soil moisture and groundwater storage capacity?" to "Where is ESP skilful, in terms of individual catchment storage capacity?", and will qualify that 'storage capacity is minus snow and water in river channel'.

We have also tested using (X1 + X3) in Section 4.3 and Figure 8, instead of X1 and X2 individually. Results are shown in the below redrawn Figure 8 (left using MSESS and right using the CRPSS, as suggested by reviewer #3). First is that results are virtually the same independent if MSESS or CRPSS is used. Interestingly, the Spearman's correlation coefficient is higher against MSESS for (X1 + X3; ρ = 0.81), than for X1 (ρ = 0.73) or X3 (ρ = 0.57) individually, and is also higher against the BFI for (X1 + X3; ρ = 0.87), than for X1 (ρ = 0.76) or X3 (ρ = 0.74). Therefore, Section 4.3 and Figure 8 will be replaced with the combined catchment storage variable (X1 + X3), instead of X1/X3 individually.



New Figure 8: Redrawn using MSESS for comparison with original manuscript (left), and using the CRPSS as is proposed metric within the revised manuscript.

R#2-6. Section 4.3 aims at finding factors for skill in the model. Did the authors check if the initial states of the model show a correlation with skill? For example, the initial amount of water in the basin, S + R in Fig. 1 of Perrin et al., 2003 (production store + routing store fillings) and the initial snow pack (if a snow model is used) can give good insight (see Singla et al., 2012).

Thank you for this really interesting suggestion. We did not yet explore if initial states show a relationship with skill, but this would certainly be a fruitful avenue for further research into a more detailed attribution of the sources of ESP skill. We feel the revised Figure 8, as outlined in R#2-5, is at a suitable level of detail for the first assessment paper and will certainly pursue this research idea in more detail in our ongoing work, thank you!

Minor comments:

R#2-7. Abstract: there is a mix between present tense and past tense. Line 14: missing S at ensembleS. Also, lines 21-22 there is a mix between lower, lowest, higher and highest. It is not known from the abstract what the rho symbol represents.

Thank you for these suggestions: We will change issue on P1;L14 to "to produce **a** 51-member ensemble of streamflow hindcasts". We will ensure all tenses are consistent and these text issues are addressed. We will add Spearman's rank correlation coefficient instead of rho symbol.

R#2-8. P. 3, L. 21: Section 5 should be Sect. 5 to be consistent with the other occurrences.

Will change.

R#2-9. P. 3, L. 28: please check all fonts sizes

Will change.

R#2-10. P. 6, L. 2: initialisation is misspelled

Will change.

R#2-11. P. 6, L. 3: at p. 5, L. 21, m is the ensemble, not the ensemble size. Also, LT means lead time, it is therefore better not to use LT for designing the number of lead times

Will make consistent.

R#2-12. P. 6, L. 4: no need for volumes, I think that streamflow is enough

Based on review comments from the other two reviewers ow we refer to the aggregation of streamflow has been made more clear as per R#1-4 and R#3-3.

R#2-13. P. 6, L. 15: remove the comma after Wilks

Will change.

R#2-14. Section 4.1.1, P. 7, L. 26 and later on: do we really need such a precision for all the scores?

We agree with the reviewer that the third decimal point in the skill scores/correlations was not necessary and will edit all instances in figures and text throughout the revised manuscript.

R#2-15. P. 9, L. 6: replace “is” with “in” (I think). In this section, percentages sometimes have a space between the figure and the percent sign, sometimes not.

Yes, will change and make spacing consistent with HESS guidelines.

R#2-16. P. 9, L. 13: is “E” actually “SE”?

Yes, good spot, will change.

R#2-17. P. 12, L. 4-6: yes, that definitely has an impact in some basins!

Indeed, while we show that it is only a very small fraction of basins studied has a significant fraction of snow, and even in these cases is usually only for winter months, it is nonetheless an important consideration within ongoing work and this is acknowledged in the text.

R#2-18. Ghannam et al. reference has some misspelling in the authors’ list

Will correct.

R#2-19. Table 1 caption: I would add “R package (Coron et al., 2016, 2017)” after “airGR” and “(Perrin et al., 2003)” at the end of the caption

We will also cite these sources in the caption: “* $\overline{F_s}$ calculated using the CemaNeige snow-accounting module (Valéry et al., 2014) within the airGR package (Coron et al., 2016, 2017) applied to the GR4J model (Perrin et al., 2003)”.

R#2-20. Table 2 caption: please remind the GR4J calibration period for the parameters that are given here.

The Tabble 2 caption will read in the revised manuscript: “Summary statistics of GR4J calibrated parameters and performance metrics for the UK and nine hydroclimate regions shown in Fig. 1. The median across n catchments within each region is given with the 5^{th} and 95^{th} percentile ranges in brackets. Calibration (Cal) was over the complete period (CP, water years 1983-2014) while evaluation (Eval) for both period 1 (P1, water years 1983-1998) and period 2 (P2, 1999-2014)”.

R#2-21. Figure 3: I think that “short”, “extended”, “monthly”, “seasonal” and “annual” should indicating more precisely what they refer to. Maybe use some arrows for this.

These terms refer directly to text on Pg7; L12-13 and Figure 3 has now been redrawn as per comment R#1-14 so we believe it is less cluttered and easier to see the vertical lines these terms directly relate to. We will also make this clearer in the revised figure caption.

References:

Singla, S., Céron, J.P., Martin, E., Regimbeau, F., Déqué, M., Habets, F., Vidal, J.-P. Predictability of soil moisture and river flows over France for the spring season (2012) Hydrology and Earth System Sciences, 16 (1), pp. 201-216.

Trinh, B.N., Thielen-del Pozo, J., Thirel, G. The reduction continuous rank probability score for evaluating discharge forecasts from hydrological ensemble prediction systems
(2013) Atmospheric Science Letters, 14 (2), pp. 61-65.

We thank Guillaume Thirel again for taking the time to provide a constructive and thorough review of our manuscript, his comments will improve our revised manuscript substantially and has given us plenty of ideas for taking the hydrological modelling elements of this work forward.

Kind regards,

Shaun.